Project Overview: Managing Big Data with MySQL

Introduction
The "Managing Big Data with MySQL" project is part of the "Excel to MySQL: Analytic Techniques for Business" specialization offered by Duke University on Coursera. This project involves analyzing a complex relational database, the Dognition database, which tracks dog behavior and customer interactions. The primary objective is to demonstrate proficiency in MySQL by executing advanced SQL queries, performing data cleaning, and generating actionable business insights.

Objective
The main objective of this project is to analyze customer behavior patterns and test completion rates using MySQL. The project aims to:
- Execute complex SQL queries to analyze relationships between multiple tables.
- Implement data cleaning techniques to ensure data quality.
- Generate meaningful business insights from raw data.
- Develop reusable SQL templates for common business analysis scenarios.

Steps Performed
The project is divided into 12 labs, each focusing on different aspects of MySQL and data analysis. Here, we'll provide a detailed summary of each lab, with a focus on the last two labs.

Lab 1: Introduction to MySQL
- Learned the basics of MySQL and relational databases.
- Set up the MySQL environment and connected to the Dognition database.
- Executed simple SQL queries to retrieve data from the database.

Lab 2: Data Retrieval and Filtering
- Explored data retrieval techniques using SELECT statements.
- Implemented WHERE clauses to filter data based on specific conditions.
- Practiced using logical operators to combine multiple conditions.

Lab 3: Data Aggregation and Grouping
- Learned to use aggregate functions like COUNT, SUM, AVG, MIN, and MAX.
- Implemented GROUP BY clauses to group data and calculate aggregate values.
- Practiced using HAVING clauses to filter grouped data.

Lab 4: Advanced Data Retrieval
- Explored advanced data retrieval techniques using subqueries.
- Implemented nested queries to perform multi-step analyses.
- Practiced using derived tables to simplify complex queries.

Lab 5: Data Cleaning and Validation
- Learned data cleaning techniques to handle NULL values and duplicates.
- Implemented data validation procedures to ensure data quality.
- Practiced using UPDATE and DELETE statements to clean data.

Lab 6: Joining Tables
- Explored different types of JOIN operations (INNER, LEFT, RIGHT).
- Implemented JOINs to analyze relationships between multiple tables.
- Practiced using JOINs to combine data from different sources.

Lab 7: Advanced Joins and Subqueries

- Learned to use advanced JOIN techniques like self-joins and cross-joins.
- Implemented subqueries within JOIN operations for complex analyses.
- Practiced using correlated subqueries to filter data based on related tables.

Lab 8: Data Transformation and Calculation
- Explored data transformation techniques using SQL functions.
- Implemented calculations and transformations within SQL queries.
- Practiced using CASE statements to create conditional logic.

Lab 9: Data Analysis and Reporting
- Learned to create complex reports using SQL.
- Implemented advanced data analysis techniques to generate insights.
- Practiced using window functions to perform calculations over partitions.

Lab 10: Performance Optimization
- Explored techniques to optimize SQL query performance.
- Implemented indexing and query optimization strategies.
- Practiced using EXPLAIN statements to analyze query execution plans.

Lab 11: Testing Relationships Between Test Completion and Dog Characteristics

Assessment 1: Dognition Personality Dimensions Analysis
- Analyzed the relationship between Dognition personality dimensions and test completion totals using complex JOIN operations and aggregate functions.
- Key findings:
  - Processed data from 35,050 unique dog profiles.
  - Identified correlations between personality types and test completion rates.
  - Achieved 89.62% accuracy in personality dimension validation.

Assessment 2: Dog Breed Analysis
- Examined relationships between dog breeds and test completion rates.
- Analyzed 758 distinct dog breeds.
- Evaluated completion patterns across 9 breed groups.
- Identified highest completion rates in Herding (11.24 tests) and Sporting (10.99 tests) groups.

Assessment 3: Breed Type Impact
- Investigated the relationship between breed types and test completion.
- Key fields: dog_guid, user_guid, breed, breed_type, breed_group.
- Contains demographic and characteristic data for each dog.
- Pure Breed: 10.41 average tests completed.
- Cross Breed: 10.60 average tests completed.
- Popular Hybrid: 10.84 average tests completed.
- Mixed Breed: 10.27 average tests completed.

Assessment 4: Neutering Impact Analysis
- Analyzed the relationship between neutering status and test completion.
- Neutered dogs completed 1-2 more tests on average.
- Implemented cross-table analysis with breed information.
- Developed complex queries using multiple JOIN operations.

Lab 12: Testing Circumstances Impact Analysis

Assessment 1: Complex Data Validation
- Implemented sophisticated data cleaning and validation procedures.
- Identified and excluded 620 negative duration entries.
- Processed 193,246 completed tests.
- Developed median calculation procedures for accurate time analysis.

```sql
SELECT test_name,
    AVG(TIMESTAMPDIFF(minute, start_time, end_time)) AS AvgDuration,
    STDDEV(TIMESTAMPDIFF(minute, start_time, end_time)) AS StdDevDuration
FROM exam_answers
WHERE TIMESTAMPDIFF(minute, start_time, end_time) > 0
GROUP BY test_name;
```

Assessment 2: Advanced Statistical Analysis
- Performed comprehensive statistical analysis of test completion patterns.
- Average test duration: 11,233 minutes.
- Implemented standard deviation calculations for reliability measures.
- Created sophisticated aggregation queries for pattern identification.

```sql
SELECT d.breed_type,
    COUNT(DISTINCT d.dog_guid) as num_dogs,
    AVG(r.rating) as avg_rating
FROM dogs d
LEFT JOIN reviews r ON d.dog_guid = r.dog_guid
WHERE d.breed_type IS NOT NULL
GROUP BY d.breed_type
HAVING COUNT(r.rating) >= 10;
```

Technical Skill Demonstration
Throughout the project, several technical skills were demonstrated, including:

- **Complex JOIN Operations**: Utilized INNER, LEFT, and RIGHT JOINs to analyze relationships between multiple tables.
- **Aggregate Functions**: Implemented COUNT, SUM, and AVG functions with GROUP BY clauses to generate business insights.
- **Subqueries and Derived Tables**: Used subqueries and derived tables for multi-step analyses.
- **Data Cleaning Techniques**: Handled NULL values and duplicate entries to ensure data quality.
- **Advanced Data Filtering**: Applied WHERE clauses with multiple conditions for precise data filtering.
- **Statistical Analysis**: Calculated standard deviations and performed median calculations for accurate data analysis.

Key Outcomes
- Successfully analyzed 193K+ customer interactions.
- Achieved 89.62% accuracy in data validation.
- Identified significant patterns in test completion rates.
- Generated actionable insights for business optimization.

Conclusion

The advanced analysis conducted in Labs 11 and 12 revealed significant patterns in test completion rates across different dog characteristics and testing circumstances. The implementation of sophisticated SQL techniques enabled deep insights into user behavior and platform performance, providing valuable data for business optimization strategies.